PB80-178353

Evaluation of the Real-Time Air-Quality Model Using the RAPS (Regional Air Pollution Study) Data Base. Volume 2. Statistical Procedures

SRI International
Menlo Park, CA

Prepared for

Environmental Sciences Research Lab.
Research Triangle Park, NC

Feb 80

U.S. Department of Commerce
National Technical Information Service
NTIS

PB80-178357

United States
Environmental Protection
Agency

Environmental Sciences Research
Laboratory
Research Triangle Park NC 27711

EPA-600/4-80-013b
February 1980

Research and Development

EPA

Evaluation of the Real-Time Air-Quality Model Using the RAPS Data Base

Volume 2. Statistical Procedures

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U. S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U.S. Environmental Protection Agency, have been grouped into nine series. These nine broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields. The nine series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies
6. Scientific and Technical Assessment Reports (STAR)
7. Interagency Energy-Environment Research and Development
8. "Special" Reports
9. Miscellaneous Reports

This report has been assigned to the ENVIRONMENTAL MONITORING series. This series describes research conducted to develop new or improved methods and instrumentation for the identification and quantification of environmental pollutants at the lowest conceivably significant concentrations. It also includes studies to determine the ambient concentrations of pollutants in the environment and/or the variance of pollutants as a function of time or meteorological factors.

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/4-80-013b	2.
3. RECIPIENT'S ACCESSION NO. PB 80 178353	
4. TITLE AND SUBTITLE EVALUATION OF THE REAL-TIME AIR-QUALITY MODEL USING THE RAPS DATA BASE Volume 2. Statistical Procedures	5. REPORT DATE February 1980 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Harold S. Javitz and Ronald E. Ruff	8. PERFORMING ORGANIZATION REPORT NO. Final Report SRI Project 6868
9. PERFORMING ORGANIZATION NAME AND ADDRESS SRI International 333 Ravenswood Avenue Menlo Park, California 94025	10. PROGRAM ELEMENT NO. 1AA603 A-26 (FY-77) 11. CONTRACT/GRANT NO. 68-02-2770
12. SPONSORING AGENCY NAME AND ADDRESS Environmental Sciences Research Laboratory — RTP, NC Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, North Carolina 27711	13. TYPE OF REPORT AND PERIOD COVERED FINAL 8/77-4/79 14. SPONSORING AGENCY CODE EPA/600/09
15. SUPPLEMENTARY NOTES	

16. ABSTRACT

The theory and programming of statistical tests for evaluating the Real-Time Air-Quality Model (RAM) using the Regional Air Pollution Study (RAPS) data base are fully documented in four volumes. Moreover, the tests are generally applicable to other model evaluation problems. Volume 2 presents the tests considered for evaluating air-quality simulation models in general and those that would be useful in evaluating the RAM. The capability of the RAM to predict sulfur dioxide (SO_2) concentrations was of particular interest. Specific tests for both intermediate and final evaluations are recommended, with accompanying descriptions of formats, plots, and procedures to establish confidence limits. Discussion focuses on the mathematics, procedures, and interpretation of the individual tests; their relevance to the project objectives; and possible trade-offs among tests.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
* Air pollution * Mathematical models * Evaluation * Tests * Computer systems programs * Statistical tests	Real-Time Air-Quality Model Regional Air Pollution Study Data Base	13B 12A 14B 09B

18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS *(This Report)* UNCLASSIFIED 20. SECURITY CLASS *(This page)* UNCLASSIFIED	21. NO. OF PAGES 22. PRICE

EPA Form 2220-1 (9-73)

EPA-600/4-80-013b
February 1980

EVALUATION OF THE REAL-TIME AIR-QUALITY MODEL USING THE RAPS DATA BASE

Volume 2. Statistical Procedures

by

Harold S. Javitz and Ronald E. Ruff
Atmospheric Science Center
SRI International
Menlo Park, California 94025

Contract No. 68-02-2770

Project Officer

John S. Irwin
Meteorology and Assessment Division
Environmental Sciences Research Laboratory
Research Triangle Park, North Carolina 27711

ENVIRONMENTAL SCIENCES RESEARCH LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27711

DISCLAIMER

This report has been reviewed by the Environmental Sciences Research Laboratory, U.S. Environmental Protection Agency, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the U.S. Environmental Protection Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

ABSTRACT

The theory and programming of statistical tests for evaluating the Real-Time Air-Quality Model (RAM) using the Regional Air Pollution Study (RAPS) data base are fully documented in four volumes. Moreover, the tests are generally applicable to other model evaluation problems. Volume 2 presents the tests considered for evaluating air-quality simulation models in general and those that would be useful in evaluating the RAM. The capability of the RAM to predict sulfur dioxide (SO_2) concentrations was of particular interest. Specific tests for both intermediate and final evaluations are recommended, with accompanying descriptions of formats, plots, and procedures to establish confidence limits. Discussion focuses on the mathematics, procedures, and interpretation of the individual tests; their relevance to the project objectives; and possible trade-offs among tests. This report was submitted in fulfillment of Contract No. 68-02-2770 by SRI International under the sponsorship of the U.S. Environmental Protection Agency. This report covers a period from 1 October 1977 to 1 April 1979, and work was completed as of April 1979.

CONTENTS

Abstract iii
Figures vi
Tables vi

1. Overview 1
   Introduction 1
   Intermediate and final evaluation in model verification 1
2. Evaluation Statistics 3
   General 3
   The steps involved in selecting intermediate-evaluation statistics 3
   The steps involved in selecting final-evaluation statistics 4
   The degree of interchangeability of intermediate- and final-evaluation statistics 6
   Final-evaluation statistics 7
   Intermediate-evaluation statistics 10

References 14
Appendices

A. Common notation and conventions 15
B. Final-evaluation statistics 17
C. Confidence limits 28
D. Intermediate-evaluation statistics 30

Preceding page blank

FIGURES

Number		Page
B-1	Geographical map	25
B-2	Translated vectors	26
B-3	Average of vectors	26
B-4	Residual vectors	26
B-5	Tolerance ellipse	26
D-1	Scattergram, regression line, and bands tolerance	30
D-2	Visual assessment of correlation	33
D-3	Nonlinear scatterplot relationships	33
D-4	Sample residual versus parameter plot	40
D-5	Sample autocorrelation plot	44
D-6	Sample plot of cumulative periodogram	45
D-7	Sample plot of residuals versus time	46
D-8	Adjusted sample plot of residuals versus time	47

TABLES

Number		Page
D-1	Post-Agreement	48
D-2	Prefigurance	48

SECTION 1

OVERVIEW

INTRODUCTION

Two types of model verification are distinguished: (1) intermediate evaluation for model modification and (2) final evaluation for model comparison. These two types of evaluation are performed via evaluation statistics, and the steps required in the selection of appropriate statistics are discussed. For intermediate evaluation, these steps are fairly straightforward. For final evaluation, the investigator must consider the relative seriousness of different types of prediction errors and must specify a "loss function." Because of the uncertainty concerning the impact of a misprediction, these are difficult tasks. The particular statistics we have chosen for intermediate and final evaluation are explained briefly in the text and more thoroughly in the appendices. In particular, considerable attention is devoted to the accuracy score--a general-purpose final-evaluation statistic.

INTERMEDIATE AND FINAL EVALUATION IN MODEL VERIFICATION

Verification of an air-pollution model is the process of comparing predicted and observed readings to evaluate model performance. The evaluation may be used for (1) modifying the air-pollution model or (2) quantifying its relative or absolute accuracy. In the first case (which we shall term "intermediate" evaluation), our test might show that the model tends to overpredict pollution concentrations at times when low wind speeds prevail. We may, therefore, choose to modify the model to compensate for the overprediction. In the second case (which we shall term "final" evaluation), we may demonstrate that the average error in prediction of a particular model is only 18 percent of the observed concentration, and that this is a lower average error than any competing model.

Intermediate evaluation is often the basis for model modification and precedes the final evaluation of the "best" model. The terminology assumes there are several models (or versions of the same model). The "best" model is that one which most accurately represents the physical processes occurring in the atmosphere as determined by comparison of the final evaluation statistic for each model. Thus, final evaluation is the basis for model comparison and precedes the selection of the best model.

During the intermediate evaluation phase, we are generally "fine tuning" a model after analyzing results from our statistical tests applied to a test data base. Hence, we run the risk of our results being unique to the test data base. To ensure that we are in fact improving a model, we must expand our evaluation using other data bases. (Alternatively, we may use a subset of a data base for intermediate evaluation and a separate, statistically independent, subset for final evaluation.)

SECTION 2

EVALUATION STATISTICS

GENERAL

Evaluation statistics are indices or scores computed from the data base of observed and predicted concentrations. These statistics constitute the method by which the evaluation is performed. For example, an intermediate evaluation of a model may involve the computation of many correlation coefficients between observed and predicted concentrations under different meteorological conditions. Certain types of statistics are better suited to intermediate evaluations and others are better suited to final evaluations. (In addition, graphical displays and accompanying subjective judgments may be appropriate for intermediate evaluations.) There are no hard and fast lines, however, and a statistic may be applied to both types of evaluations.

THE STEPS INVOLVED IN SELECTING INTERMEDIATE-EVALUATION STATISTICS

Different steps are involved in selecting the appropriate statistic to accomplish an intermediate or a final evaluation.

For an intermediate evaluation, we first specify an aspect of the agreement (or disagreement) between the predicted and observed concentrations we intend to investigate. For example, we may be interested in investigating whether high (low) predicted concentrations are associated with high (low) observed concentrations. Or we may be interested in investigating cyclical patterns of overprediction and underprediction. Secondly, we translate the aspect of the agreement we wish to investigate into a model for agreement. Such a model for agreement might be that observed and predicted concentrations are random variables following a bivariate normal distribution with an unknown correlation coefficient. Alternatively, we might postulate that the error in prediction is composed of a sinusoidal component and a white-noise component. Thirdly, we

specify a statistic that is optimal in some sense for estimating or testing the postulated model or some aspect thereof. For example, Pearson's sample correlation coefficient is often used to estimate the true correlation coefficient. Or the normalized cumulative periodogram may be used to test for the presence of cyclical trends in prediction error. (Both are described in more detail later in this Section.)

THE STEPS INVOLVED IN SELECTING FINAL-EVALUATION STATISTICS

The Relative Seriousness of Errors in Prediction

For the final evaluation, the first step is determining the relative seriousness (or "loss") associated with different types of errors in air pollution prediction. Consideration should be given to the following types of questions:

- Is an underprediction more or less serious than an overprediction?
- Is the assessment of the seriousness of an underprediction or an overprediction also dependent upon the magnitude of the observed concentration?
- Can misprediction errors be categorized according to quantifiable losses (such as the costs incurred in calling erroneous air pollution alerts)?
- How serious is an error in predicting the location of the maximum air pollution?
- Is a string of overpredictions that is followed by a string of underpredictions more or less serious than a more interspersed pattern?

Specifying a Loss Function

The end result of these considerations is the specification of a loss function. In its most general form, the loss function quantitatively determines, for any set of observed and predicted concentrations, the loss associated with any errors in prediction. The data base to which the loss function is applied consists of the observed and predicted concentrations at one or more spatial locations and at one or more time periods. The data base may also include meteorological and other types of data that are relevant to the computation of loss. Consequently, the loss function may be quite complex. The evaluation statistic associated with an

air-pollution model and a data set is simply the loss associated with the data set. For example, a relatively simple evaluation statistic would be the average-squared error between observed and predicted concentrations over n time periods at a single location or group of locations. Mathematically, this statistic may be written

$$\frac{1}{n}\sum_{i=1}^{n}(OC_i - PC_i)^2$$

where OC_i is the observed concentration at the ith time period and PC_i is the corresponding predicted concentration.* Using this statistic implies that the investigator considers underprediction and overprediction to be equally serious, and the assessment of the seriousness of an error in prediction to be independent of the magnitude of the observed concentration, the existence of any threshold, the location of the maximum predicted or observed concentration, or any temporal arrangement of the underpredictions and overpredictions.

The Difficulty in Specifying a Loss Function

Specifying a final-evaluation statistic is equivalent to specifying the loss function that quantifies the losses associated with every type of error in prediction. While generally acceptable final-evaluation statistics (such as mean square error) have been used in the past, the associated loss function and its implications have not been uniformly considered. This omission appears to be linked to the difficulty in quantifying the impacts of misprediction.

As an evaluation tool in the model-development phase, the loss function must be objective so that we can evaluate changes to a specific model or make comparisons among competing models. However, when a model is applied in an operational mode, many subjectively derived factors may

---

*Refer to Appendix A for a complete description of the notation used throughout this text.

figure in the evaluation process. Difficulties are encountered both in the quantification of the utility or lack thereof to each user of a prediction and in the aggregations of these over many users. Consider the following examples:

- A pollution alert may convince an individual to refrain from any vigorous physical activity for a day or longer, with the individual incurring both an opportunity loss and a possible economic loss, balanced against a projected medical gain.
- A pollution alert might result in the voluntary or mandatory shutdown of local factories.
- The failure to call an alert might result in an unusually large number of emphysema attacks or deaths.
- Decisions to allow or forbid the construction of factories might be partially based on prediction forecasts.

In these examples, the impacts of misprediction can be economic, medical, and political. Such impacts are difficult to quantify and aggregate.

Nevertheless, specifying the relative seriousness of different types of errors in air pollution prediction is an absolutely vital step in final evaluation.* In the absence of a prior study on relative losses, we advocate the specification of a reasonable loss function, obtained through a process for generating a consensus opinion.

THE DEGREE OF INTERCHANGEABILITY OF INTERMEDIATE- AND FINAL-EVALUATION STATISTICS

As mentioned earlier, although a particular statistic may be particularly well suited for intermediate or final evaluation, there is no logical prohibition against using it for the other evaluation purpose.

---

*An "alternative" to the specification of a single final-evaluation statistic is the specification of a battery of statistics, some of which are appropriate to final evaluation and some to intermediate evaluation. Investigators use the battery of statistics to gain an appreciation of the strengths and weaknesses of the model. They then subjectively weight the relative importance of these strengths and weaknesses in order to compare models. This type of final evaluation may be useful, but it is less rigorous, scientific, objective, or clearly specified than the type of final evaluation we have advocated.

The average-squared-error statistic just discussed is an example of a statistic that may be profitably used in an intermediate evaluation. In general, we find that the final-evaluation statistics associated with simpler loss functions are also useful for intermediate evaluation; however, statistics appropriate to intermediate evaluation appear in general to be rather poorly suited to final evaluation.

A particularly glaring example of this phenomenon is Pearson's correlation coefficient (which, unfortunately, is often used in final evaluations). When performing a final evaluation, it is reasonable to implicitly associate a loss of zero with a Pearson's correlation coefficient of +1 and a large loss with a Pearson's correlation coefficient of -1. This loss function necessitates some very strange considerations when final evaluations are chosen. For example, if Model A predicts the observed concentrations very accurately and Model B always predicts ten times the concentrations of Model A, then both models have the same Pearson's correlation coefficient. Again, if Model C always predicts 50 ppb more pollution than Model A, then both these models have the same Pearson's correlation coefficient. The investigator would presumably find all these models equally acceptable. The near-irrationality of these acceptability conditions highlights the unsuitability of using Pearson's correlation coefficient as a final single-evaluation statistic.*

FINAL-EVALUATION STATISTICS

The Accuracy Score--A General Final-Evaluation Statistic

We have formalized a number of reasonable loss functions and their associated statistics, which we have termed accuracy scores. Accuracy

---

*If, in the intermediate-evaluation phase, all prediction models are adjusted so that the average predicted concentration is the same as the average observed concentration, and the variance of the predicted concentration is the same as the variance of the observed concentration, then Pearson's correlation coefficient is mathematically equivalent to mean-square error. In this event, either Pearson's correlation coefficient or mean-square error may be used as a final-evaluation statistic, with equivalent results.

scores are based on a particularly simple, but quite general, type of loss function. These are the loss functions that are additive over time periods. A loss function is additive over time periods if (1) a loss can be computed due to the error in prediction at one particular time period and (2) the total loss due to error in prediction over a set of time periods is the arithmetic sum of the losses in each time period in the set. The accuracy score is the average loss over the set of time periods. An example of an accuracy score is average-squared loss, which has already been discussed. Eight different accuracy scores are described in the first section of Appendix B. Seven of these are concerned with the losses due to errors in prediction at a particular location. If data are available for many locations, one may base the final evaluation on the average loss over the total number of prediction-observation pairs.* The eighth accuracy score is concerned with losses due to an error in predicting the location of the maximum observed concentration.

More complicated accuracy scores are also possible. For example, one might specify an accuracy score based on the loss for a particular time period while simultaneously considering both the error in the prediction of the maximum concentration and the error in the prediction of the location of the maximum concentration. In addition, one might also base final evaluation on a statistic other than an accuracy score (for example, a statistic based on a nonadditive loss function†) if this better reflected the concept of relative losses.

---

*Alternatively, one could base the final evaluation on the largest value of the accuracy score when a particular accuracy score is computed over many locations. This final-evaluation statistic should not be used without careful thought. For example, it implies that if Model A predicts perfectly accurately at Location 1 and poorly at Location 2, and Model B predicts, at Locations 1 and 2, as poorly as Model A does at Location 2, then Models A and B are equally desirable.

†A nonadditive loss function might be appropriate in certain situations. For example, suppose that an alert is called only after three consecutive hourly predictions of excessive pollution for the next 24-hour period. In this case, very little loss would be associated with overprediction for a single time period, since it would take three consecutive overpredictions to trigger a false alert.

To supplement the accuracy score, we recommend two features that enhance its meaning and interpretation. First, confidence limits can be computed so that its statistical significance can be assessed. (Refer to Appendix C for a description of the general procedure.) As a second feature, the accuracy score for each observation point (monitoring location) can be plotted on a geographical map of the area. The resulting display can then be analyzed for geographical trends. For instance, we could determine if the higher accuracy scores corresponded to urban areas or rural areas, or if the highest scores were associated with one geographical sector (e.g., northwest). Thus, such a display could serve to enhance the accuracy score's overall utility.

Other Final-Evaluation Statistics

Prior to arriving at our recommendation of the accuracy score for evaluating the final model, we considered several other tests. Each has its own advantages. These other tests were considered in detail but not included in our final package because they repeated, to a large extent, the accuracy-score statistics. The three tests (skill score, P-score, and maximum location tolerance intervals) are described below and in Appendix B.

Skill Score--

The skill score discussed in Appendix B (see p. 21), is a statistic that can be used either as an intermediate- or as a final-evaluation statistic. The skill score is an index of model performance, which may be adjusted to reflect how well a baseline model (such as chance or persistence) would have performed. This statistic is substantially the same as the accuracy score, with a user-supplied loss matrix (see the write-up of the accuracy score in Appendix B) with zeros on the diagonal and a loss of 1 elsewhere.

P-Score--

The P-score statistic, as described in Appendix B (see p. 22), is useful in evaluating models that assign probabilities to various levels

of concentration. This type of prediction is not common, however, and the accuracy score could be adjusted to accommodate this type of model.

Maximum Concentration Tolerance Intervals--

The purpose of this test is to evaluate a model's ability to predict the location of the maximum (observed) concentration. (See p. 23 of Appendix B). This test computes the mean distance between the "observed" and predicted locations of maximum concentration and a spatial tolerance interval (which would basically be defined by an ellipsoid). The drawback with this test is that the monitoring network must be of sufficient spatial density that the observed concentration is known everywhere in the region of interest.

Realistically, we must expand our number of observation locations through interpolation techniques. If the monitoring network is not dense enough to support an accurate interpolation, the test is neither objective nor statistically sound. For the RAPS network, monitoring networks are separated by about 5 km, and the ability to interpolate air-quality concentrations between observational stations is largely unknown. Therefore, we have not recommended the inclusion of this test in our statistical package. Instead, we have incorporated a test with a similar objective into our accuracy score, using a maximum concentration location loss function. However, when (or if) it is proven that accurate interpolation between monitoring sites can be achieved, inclusion of the maximum concentration tolerance intervals should be reconsidered.

INTERMEDIATE-EVALUATION STATISTICS

In addition to the recommended final-evaluation statistical tests contained in the accuracy score, we are recommending several tests that are useful in diagnosing problem areas within models. Results from these "intermediate-evaluation" tests serve to identify components of the model most in need of improvement. In some cases, these tests can serve as final-evaluation criteria as well. The following tests and displays are recommended for the above type of intermediate evaluation:

- Scatterplot
- Linear regression with confidence and prediction bands
- Pearson's correlation coefficient
- Chi-square test and frequency distribution displays
- Interstation error correlation
- Multiple regression of error residuals
- Analysis of residual time series.

The above and other tests considered in our study are discussed below and in Appendix D.

Discussion of the Tests Selected

Perhaps the most basic and most useful intermediate-evaluation technique is the graphical display of the data in a scatterplot. The evaluation of the scatterplot of predicted concentration (on the vertical axis) versus the observed concentration (on the horizontal axis) is aided by calculating the linear regression line and the sample variance around this line. The linear regression line may be used directly to modify the model, so that the sample variance around the line becomes the new root-mean-squared error. These techniques are discussed in Appendix D (see p. 29). This section also discusses the confidence bands that can be placed around the linear regression line and the prediction bands for additional observations. The width of the first of these bands indicates how well the position of the "true" underlying linear regression is known from the sample (if indeed the relationship is linear), while the width of the second band indicates how far from the linear regression a new observation could be expected to fall.

Closely allied to linear regression is Pearson's correlation coefficient, which has long been used as a model evaluation statistic. This statistic is described in Appendix D (see p. 31), along with the confidence bands for the correlation coefficient.

In Appendix D (p. 34), the spatial and temporal correlation coefficients are presented along with their appropriate confidence bands. The spatial correlation coefficient measures the agreement between the

predicted and observed spatial trends of air pollution over a prediction period. The temporal correlation coefficient measures the agreement of the predicted and observed temporal trends of air pollution over the whole network of monitoring stations. These statistics are essentially derivatives of Pearson's correlation coefficient and contain much the same information.

The chi-square test is a standard statistic used for examining the fit of the distribution of predicted concentrations to the distribution of observed concentrations. This statistic and some useful accompanying graphics are discussed in Appendix D (see p. 36). While useful in general, this statistic does not examine the observed and predicted concentrations pairwise. For this reason we suggest that it be used only as a back-up statistic for a procedure such as linear regression, which does examine the data pairwise. The chi-square test is also a useful final-evaluation statistic for those cases where the accuracy of the predicted concentration distribution is important.

A number of techniques can be used to examine directly the error residuals (that is, the differency between the observed and predicted concentrations). Appendix D (p. 38) introduces the interstation error correlation coefficient--a statistic that examines the correlation between station errors, which may be helpful in uncovering influences accounting for systematic biases in the prediction model. Another, more direct, method is the multiple regression of error residuals, also discussed in Appendix D (see p. 39). This technique attempts to isolate meteorological variables highly correlated with the errors in prediction. Finally, the analysis of residual time series discussed in Appendix D (p. 41) examines the error residuals for any cyclical behavior using two types of statistics--the autocorrelation coefficient and the normalized cumulative periodogram.

Other Tests

Appendix D also includes four statistics seriously considered by SRI as evaluation statistics but finally rejected. The runs test (see p. 45) is a nonparametric test that examines for time trends

in error residuals. The analysis of residual time series accomplishes more than the runs test, however, and is more easily interpreted. The prefigurance and post-agreement tables (see p. 46) allow an assessment of the extent to which the forecasts give advance notice of a certain pollution level and the extent to which subsequent observations confirm this prediction. However, the scatterplot contains the same information in a more rapidly digestible form, without the inconvenience of the user specifying pollution concentration categories.

Quasi-logistic modeling of the correlation coefficient (see p. 48) is a statistical technique for determining the meteorological conditions under which the correlation coefficient is unusually high or low. The data set must be quite extensive before this technique can be used, however, and the multiple regression of error residuals can extract similar information more efficiently.

REFERENCES

1. R. G. Miller, Simultaneous Statistical Inference, pp. 110-116 (McGraw-Hill, New York, New York, 1967).

2. T. W. Anderson, An Introduction to Multivariate Statistical Analysis, pp. 77-79 (John Wiley and Sons, New York, New York, 1958).

3. R. Wonnacott and T. Wonnacott, Econometrics, pp. 103-125 (John Wiley and Sons, New York, New York, 1970).

4. Carmen J. Nappo, "A Method for Evaluating the Accuracy of Air Pollution Prediction Models," Proc. Conf. on Atmospheric Diffusion and Air Pollution, 9-13 September 1974, Santa Barbara, California.

5. Robert Hogg and Allen Craig, Introduction to Mathematical Statistics, pp. 308-315 (The Macmillan Company, New York, New York, 1970).

6. N. R. Draper and H. Smith, Applied Regression Analysis, pp. 104-195 (John Wiley and Sons, New York, New York, 1966).

7. George E. Box and Gwilym M. Jenkins, Time Series Analysis Forecasting and Control, pp. 34-36 and 294-298 (Holden-Day, San Francisco, California, 1976).

8. Gwilym M. Jenkins and Donald G. Watts, Spectral Analysis and Its Applications, pp. 183-189 and 234-237 (Holden-Day, San Francisco, California, 1968).

9. W. J. Conover, Practical Nonparametric Statistics, pp. 350-356 and 414 (John Wiley and Sons, New York, New York, 1971).

10. H. A. Panofsky and G. W. Brier, Some Applications of Statistics to Meteorology, pp. 191-208 (Pennsylvania State University Press, University Park, Pennsylvania, 1965).

11. Norman Nie et al., Statistical Package for the Social Sciences, pp. 320-328 (McGraw-Hill, New York, New York, 1975).

APPENDIX A

COMMON NOTATION AND CONVENTIONS

To shorten the descriptions of the statistical tests, we have assembled a common set of notation and conventions. Any exceptions to this common set are specifically mentioned in the descriptions of particular statistical tests.

In simplest form, the data for a statistical test consist of n pairs of observed and predicted pollutant concentrations. These data are denoted (OC_1, PC_1), (OC_2, PC_2), . . . (OC_n, PC_n). This data set may include all the data gathered or only a subset. In general, a final-evaluation statistic will make use of all the available data, and intermediate-evaluation statistics will most often make use of subsets. The concentrations might refer to hourly averages, three-hour averages, daily averages, etc., at a particular station or averaged over the station network. The pairs need not be sequential nor time-ordered in any particular fashion.

Since it is desirable to assess model fit under a variety of meteorological conditions, the data base is arranged so that it can be subdivided into many nonexclusive subsets according to the following parameters:

- Date, D
- Time period, T
- Location, L
- Pasquill-Gifford stability class, G_1
- Wind-speed class, G_2
- Wind-direction class, G_3
- Mixing-height class, G_4
- Emission class, Q
- Observed concentration, OC
- Predicted concentration, PC.

The data for a statistical test shall then be selected from any subset of the data base (in order to assess model fit under those conditions characterizing that data subset).

In quantifying confidence intervals, we use Student's t-distribution.[1] For large sample sizes, the t-distribution approximates that represented by the standard normal curve. In the notation, $T_{m,\alpha}$ represents the upper $\alpha/2$ cutoff point for a t-distribution with m-degrees of freedom.

[1]Selby, S. M., CRC Standard Mathematical Tables (Chemical Rubber Company, Cleveland, 1968).

APPENDIX B

FINAL-EVALUATION STATISTICS

ACCURACY SCORE

The accuracy score offers the user several tests to assess either objectively or subjectively the "loss" due to differences in predicted concentrations (PC) and observed concentrations (OC). Each test is based on the quantification of a loss function, which measures the loss associated with an error in prediction at a particular time period. The accuracy score (denoted A) is the average loss for the prediction model computed over the n time periods in the data base. Consequently, the accuracy score resembles a golf score--the lower, the better. (The text by Ferguson[1] provides material complementary to the concepts presented herein.)

We consider eight loss functions, which naturally are associated with eight different accuracy scores. The first five rely on (essentially) prespecified loss functions, which are easily interpretable and have therefore been called objective. The next two require that the user specify loss indices associated with various types of misprediction and have therefore been called subjective. The eighth loss function provides an index of the spatial accuracy of the model in predicting the location of the maximum observed concentration.

The importance of correctly specifying the loss function cannot be overemphasized. The loss function embodies our conception of how poorly or well the model fits and the relative seriousness of different departures from perfect fit. For example, a meteorologist who is concerned only about false pollution alerts will choose a very different loss

---

[1]Ferguson, Thomas S., Mathematical Statistics, pp. 1-11 (Academic Press, New York, 1967).

function than a meteorologist who is concerned both about false pollution alerts and failure to predict actual alert-level pollution.

We now describe eight loss functions and accuracy scores--five objective and three subjective.

Objective Loss Functions

Average Absolute Error--

$$A = \frac{1}{n} \sum_{i=1}^{n} |OC_i - PC_i|$$

With this loss function (absolute error), underprediction and overprediction are judged to be equally undesirable. Furthermore, a misprediction of 2 ppm is twice as undesirable as a misprediction of 1 ppm.

Average-Squared Error--

$$A = \frac{1}{n} \sum_{i=1}^{n} (OC_i - PC_i)^2$$

With this loss function (squared error) also, underprediction and overprediction are judged to be equally undesirable. A misprediction of 2 ppm is four times as undesirable as a misprediction of 1 ppm. This type of loss function is used in many statistical tests.

Percent Incorrect Predictions with Absolute Error Threshold--

$$A = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} 0 & \text{if } |OC_i - PC_i| \leq \epsilon \\ 1 & \text{if } |OC_i - PC_i| > \epsilon \end{cases}$$

With this loss function, only prediction errors that exceed a certain absolute threshold are counted against the prediction model.

Underprediction and overprediction are judged to be equally undesirable. Whenever the misprediction exceeds the threshold, its magnitude is ignored. Thus, moderate and huge misprediction errors count equally. (The error threshold must be user-supplied.)

Percent Incorrect Prediction
with Constant Percentage Error Threshold (ϵ)--

$$A = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} 0 & \text{if } \left| \frac{OC_i - PC_i}{OC_i} \right| \leq \epsilon \\ 1 & \text{if } \left| \frac{OC_i - PC_i}{OC_i} \right| > \epsilon \end{cases}$$

This loss function is similar to the preceding, except that the threshold criterion is percentage error. (In practice, the case where OC = 0 is not allowed.)

Symmetric High-Low Loss Function (Cutoff)--

$$A = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} 0 & \text{if } PC_i \text{ and } OC_i \leq \text{cutoff} \\ & \text{or} \\ & PC_i \text{ and } OC_i > \text{cutoff} \\ 1 & \text{if otherwise} \end{cases}$$

This loss function is particularly appropriate when an alert level cutoff exists. Here A is the percentage of false alerts and failures to predict actual pollutant alert levels. The magnitude of the false alert is ignored as is the magnitude of the underprediction when an actual pollutant alert condition is not predicted.

Subjective Loss Functions

Asymmetric High-Low Loss Function (Cutoff)--

$$A = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} 0 \text{ if or} & PC_i \text{ and } OC_i \leq \text{cutoff} \\ & PC_i \text{ and } OC_i > \text{cutoff} \\ \ell_1 \text{ if} & PC_i > \text{cutoff and } OC_i \leq \text{cutoff} \\ \ell_2 \text{ if} & PC_i \leq \text{cutoff and } OC_i > \text{cutoff} \end{cases}$$

Here, ℓ_1, ℓ_2, and the cutoff value are user-specified loss quantities. This loss function is similar to the preceding one, except that underprediction and overprediction are weighted differently.

User-Supplied Loss Matrix--

The predicted and observed concentrations are divided into categories whose increments are about 20 percent of the applicable standard. The user enters the loss numbers in the matrix. An example of the loss table is shown below:

PC Categories (j) \ OC Categories (k)	1	2	3	4	5
1	0	1	2	3	4
2	1	0	1	2	3
3	2	1	0	1	2
4	3	2	1	0	1
5	4	3	2	1	0

Let ℓ_{jk} be the number in the jth row and the kth column. The accuracy score is

$$A = \frac{1}{n} \sum_{i=1}^{n} \ell_{jk}$$

where PC_i is in the jth category and OC_i is in the kth category.

The loss function, ℓ_{jk}, is extremely general, since the number of cells and the loss numbers in the cells are at the user's discretion. The sample loss matrix approximates the average absolute error loss function. The other loss functions could be similarly approximated.

Loss Function for the Location of Maximum Concentration--

The previous concepts of the accuracy score can be slightly modified to provide an index of the spatial accuracy of the prediction model in forecasting the location of the maximum-concentration reading. This accuracy score can be defined in the following fashion:

Step 1: For the ith time period, identify the monitoring locations, $k(i)$ and $k'(i)$, where the highest observed and predicted concentrations occur.

Step 2: From the symmetric $m \times m$ matrix, consisting of the distances between monitoring stations, select the distance $d_{k(i),k'(i)}$ between the locations for highest observed and predicted concentrations.

Step 3: Repeat Steps 1 and 2 for all samples and then compute the accuracy score from the formula

$$A = \frac{1}{n} \sum_{i=1}^{n} d_{k(i),k'(i)} \quad .$$

Many other types of objective or subjective loss indices $\ell_{k(i),k'(i)}$ could be used instead of geographical distance. For instance, the loss index could be constructed to take into account topographical features

that might cause two geographically close locations to differ substantially in pollutant levels.

SKILL SCORE

The skill score[2] is an index of model performance that partially summarizes the information contained in the contingency table of predicted and observed pollution concentrations. The skill score may be adjusted to reflect how well a baseline model (chance, persistence, climatology, etc.) would have performed.

The skill score S is given by

$$S = \frac{C - E}{T - E}$$

where C is the number of correct forecasts and T is the total number of forecasts. If no adjustment for a baseline model is desired, then E is set to zero and the skill score is just the percentage of correct forecasts. If an adjustment for a baseline model is desired, then E is set to the expected number of correct forecasts for the baseline model. In this case, the skill score (1) has a value of unity when all forecasts are correct, (2) has a value of zero when the number of correct forecasts equals the expected number correct, and (3) is negative when the number of correct forecasts is less than the expected number correct.

The skill score may easily be modified to summarize how well a model predicts the location of maximum pollution concentration. To this end, C is defined as the number of correct location forecasts and T is the total number of location forecasts. If an adjustment for a baseline model is desired, then E is set to the expected number of correct location forecasts for the baseline model.

---

[2]Panofsky, H. A., and G. W. Brier, Some Applications of Statistics to Meteorology, pp. 191-208 (Pennsylvania State University Press, University Park, Pennsylvania, 1965).

P-SCORE

The P-score[3] is a method for evaluating a forecast model and is particularly well suited for models that predict the probabilities of certain pollutant levels occurring. This score has the desirable property that it exerts little influence on the model builder to choose a particular model in order to score well on the statistical evaluation. The forecaster is encouraged to get the forecast exactly right, or, if he cannot forecast perfectly, to state unbiased estimates of the probability of each possible event.

The P-score ranges from zero to one. A score of zero implies that the prediction model always forecasts the observed pollutant level with a probability of one. A score of one implies that the prediction model always forecasts a probability of zero for the pollutant level that did materialize and forecasts a probability of one for a pollutant level that did not materialize.

To compute the P-score, the user must specify various levels for pollutant concentration, say K of them. For the ith time period, the model is assumed to forecast probabilities that the observed concentration will be at particular levels. Let f_{i1} be the forecast probability that the observed concentration at time i will be at level 1; let f_{i2} be the forecast probability that the observed concentration at time i will be at level 2, etc. Clearly, $f_{i1} + f_{i2} + \ldots + f_{iK}$ equals one. The P-score is given by

$$P = \frac{1}{2n} \sum_{j=1}^{K} \sum_{i=1}^{n} (f_{ij} - E_{ij})^2$$

where there are n time periods, and where E_{ij} takes the value zero or one according to whether or not level j was observed at time i.

---

[3]Morrison, D. F., Multivariate Statistical Methods, pp. 131-134 (McGraw-Hill, New York, 1976).

In the absence of any forecasting skill, the expected P score is minimized by choosing f_{ij} to be the climatological probability of observing level j. If these probabilities are denoted p_j, then the minimum expected P-score is

$$1 - \sum_{j=1}^{K} p_j^2 \quad .$$

The climatological probabilities may be computed either from the sample or from the larger set of data it represents.

TOLERANCE INTERVALS FOR THE LOCATION OF MAXIMUM CONCENTRATION

The ability to predict the location of the maximum concentration of pollutants is an important characteristic of a prediction model. In this section we derive a tolerance interval (or ellipsoid) to describe this ability.

Consider a geographical map of a region. One may locate on this map the monitoring stations and the predicted location of the maximum pollution. Using the results obtained at the monitoring stations, one may determine an interpolated location for the observed maximum concentrations.

In Figure B-1 an arrow has been drawn from the location of observed maximum concentration to the location of predicted maximum concentration. Two techniques are now available to describe the ability of the model to predict the location of the maximum concentration.

The first technique is concerned only with the length of the arrow (i.e., the distance between the observed and predicted locations). Letting D_i be the distance for the $\underline{i}$th time period, we compute $\overline{D}$ and S_d, the usual estimates of mean and standard deviations. A confidence interval for U_d, the mean of the distribution of D_i, is given by

$$\overline{D} \pm t_{.95} S_D/\sqrt{n}$$

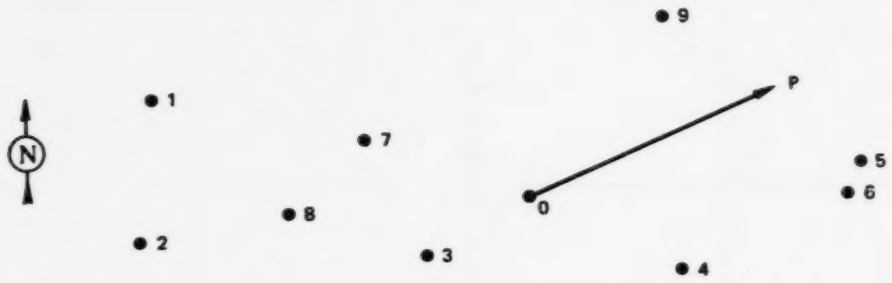

Figure B-1. Geographical map.

where $t_{.95}$ is the 95th percentile point of the t distribution, with n-1 degrees of freedom. Under the assumption that the D_i follow a normal distribution, a 90-percent tolerance interval can be developed. Consider an additional observation, D_o, not used in the original calculations. We are 90-percent confident that D_o will be in the interval:

$$\bar{D} \pm t_{.95} \left(1 + \frac{1}{n}\right)^{1/2} S_D \quad .$$

The second technique is concerned both with the distance between the observed and predicted locations and with the direction of the error. For each time period, an arrow may be drawn. Translating these arrows so that the bases are located at an origin, we obtain Figure B-2. It is easily seen that most of the predictions are to the northeast of the observations, fewer are to the southwest, and still fewer to the northwest or southeast.

Taking the average of the vectors, we find the average direction and magnitude of the error in prediction, as shown in Figure B-3. The residual vectors (from the average vector to the translated vectors) may also be drawn, as shown in Figure B-4.

Using statistical techniques, we may draw an ellipsoid, which captures most of the residual vectors, around the average vector. (See Figure B-5.) The ellipsoid is large enough that the investigator can be

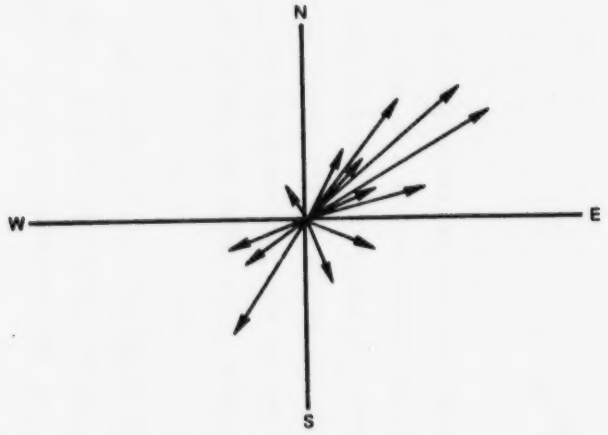

Figure B-2. Translated vectors.

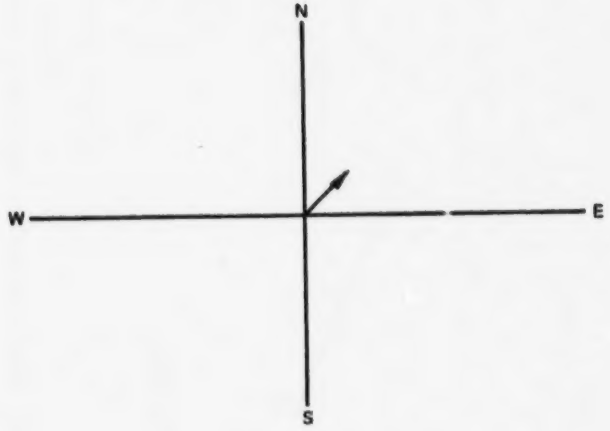

Figure B-3. Average of vectors.

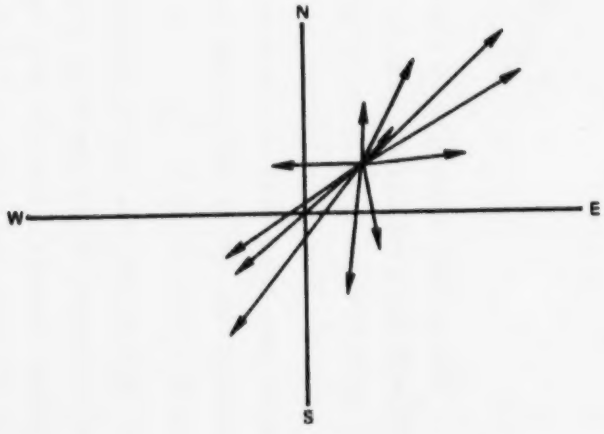

Figure B-4. Residual vectors.

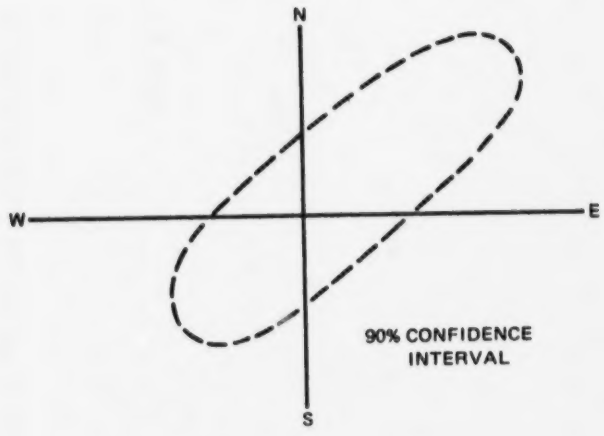

Figure B-5. Tolerance ellipse.

90-percent confident it will capture a translated vector, when such a vector is obtained under conditions similar to those previously existing. A description of the general procedure follows.

Let X denote the east-west axis and Y denote the north-south axis. Then the vector of differences between observed and predicted locations is

$$(X_i, Y_i) = \bar{D}_i \quad .$$

We assume that the vector (X,Y) follows a bivariate normal probability distribution with mean (U_x, U_y), variances (σ_x^2, σ_y^2), and correlation ρ.

The best estimate of the mean (U_x, U_y) is the sample mean $(\bar{X}, \bar{Y})$. A confidence interval for (U_x, U_y) is

$$\begin{pmatrix} U_x - \bar{X} \\ U_y - \bar{Y} \end{pmatrix}^1 \begin{pmatrix} \frac{S_x^2}{n} & \frac{S_{xy}}{n} \\ \frac{S_{xy}}{n} & \frac{S_y^2}{n} \end{pmatrix}^{-1} \begin{pmatrix} U_x - \bar{X} \\ U_y - \bar{Y} \end{pmatrix} \leq T_o^2$$

where S_x^2 and S_y^2 are the usual estimates of the variance of X and Y, and S_{xy} is the estimate of the covariance of X and Y. Here T_o^2 is the cut-off value of Hotellings' T^2 statistic.[3]

Consider an additional vector of differences, not used in the original calculations, say (X_o, Y_o). Assuming that (X_o, Y_o) follows the same bivariate normal distribution as (X,Y), we obtain the following result: we are 90-percent confident that (X_o, Y_o) will lie in the ellipsoid

$$\begin{pmatrix} X_o - \bar{X} \\ Y_o - \bar{Y} \end{pmatrix}^1 \begin{bmatrix} \left(1 + \frac{1}{n}\right) S_x^2 & \left(1 + \frac{1}{n}\right) S_{xy} \\ \left(1 + \frac{1}{n}\right) S_{xy} & \left(1 + \frac{1}{n}\right) S_y^2 \end{bmatrix}^{-1} \begin{pmatrix} X_o - \bar{X} \\ Y_o - \bar{Y} \end{pmatrix}^1 \leq T_o^2 \quad .$$

[3]Morrison, D. F., Multivariate Statistical Methods, pp. 131-134 (McGraw-Hill, New York, 1976).

APPENDIX C

CONFIDENCE LIMITS

The accuracy score and a number of the intermediate-evaluation statistics presented in this report do not have formulas for confidence limits. However, confidence limits for these statistics may be derived.

To illustrate the procedure for the accuracy score, let A be an accuracy score computed from a data set. Subdivide the data set into k data subsets. We recommend k = 5 as a reasonable value. In this case, the first subset should contain the data from time-periods 1, 6, 11, 16, etc. The second subset should contain data from time-periods 2, 7, 12, 17, etc. In this fashion each subset contains data from every fifth time period.*

The accuracy scores for each of the subsets can be computed if these accuracy scores are labeled A_1, A_2, ..., A_k; then the sample standard deviation of these subset scores is

$$S = \sqrt{\frac{1}{k - 1} \sum_{i=1}^{k} (A_i - \bar{A})^2}$$

where

$$\bar{A} = \frac{1}{k} \sum_{i=1}^{k} A_i \quad .$$

---

*This procedure ensures that the data subsets are comparable with each other and to the full data set, except when the data set has a "periodicity" of five time periods. As an example of such periodicity, consider the case wherein the periods were days and the full data set consisted only of weekdays. Then the five data sets would not be comparable (since one would consist only of Mondays, another only of Tuesdays, etc.). In this case, one should construct six data subsets.

The accuracy-score A computed from the total data set has a smaller standard error than the accuracy scores computed from the smaller data subsets. We may approximate the standard error of A by

$$\frac{S}{\sqrt{k}} \quad .$$

Thus a 90-percent confidence interval for the true value of the accuracy score is given by

$$A \pm t_{k-1,.95} \frac{S}{\sqrt{k}}$$

where $t_{k-1,.95}$ is the upper 95-percent cutoff point of the Student's t-distribution with k - 1 (four) degrees of freedom.

APPENDIX D

INTERMEDIATE-EVALUATION STATISTICS

LINEAR REGRESSION WITH CONFIDENCE AND PREDICTION BANDS

The scattergram of observed-versus-predicted concentrations is a visual presentation of the accuracy of the prediction model. The regression line summarizes the average relationship between predicted and observed concentrations.

Two types of bands can be drawn around the regression line. A 90-percent confidence band has a 90-percent probability of capturing a true underlying linear relationship between predictions and observations. A 90-percent prediction band has a 90-percent probability of capturing an arbitrary prediction-observation pair, when that pair is obtained under conditions similar to those existing when the data used in the scattergram were gathered. Typical bands are shown in Figure D-1.

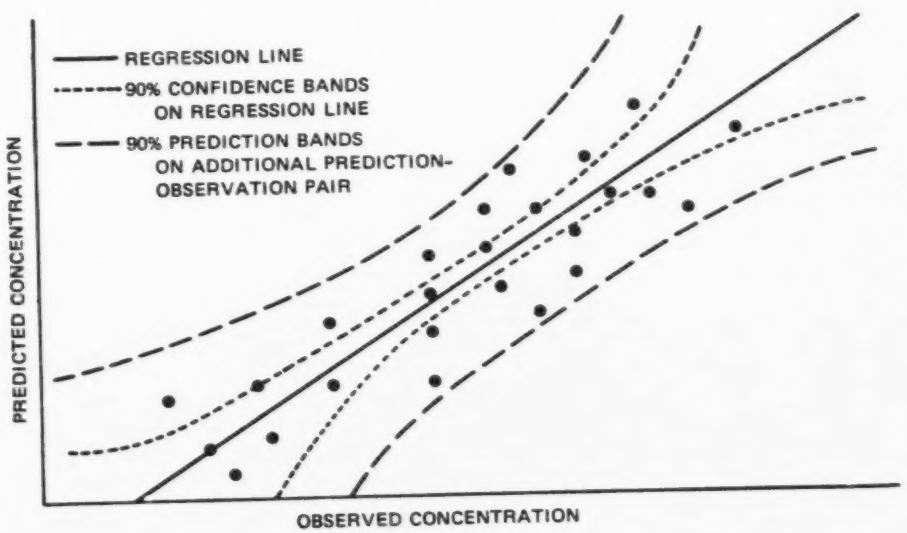

Figure D-1. Scattergram, regression line, and bands tolerance.

The regression line and bands can be obtained by applying normal regression theory. For the sake of notation, let

$$(X_i, Y_i) = (OC_i, PC_i)$$

Then the least-square regression of Y_i on X_i for n samples may be performed, yielding an intercept estimate $\hat{a}$, a slope estimate $\hat{b}$, and a variance estimate S^2 as follows:

$$\hat{a} = \frac{\left(\sum_{i=1}^{n} Y_i\right)\left(\sum_{i=1}^{n} X_i^2\right) - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} X_i Y_i\right)}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

$$\hat{b} = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i)^2$$

In contrast to the usual convention, the vertical axis has been reserved for the predicted observation and the horizontal axis has been reserved for the observed concentration. This arrangement more closely approximates the underlying structure of normal regression theory since the predominant error is in the prediction rather than in the observation of air pollution.

The usual 90-percent confidence interval for the mean $\hat{a} + \hat{b}X$ is given by

$$\hat{a} + \hat{b}X \pm t_{n-2,.95}\, S\left[\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\right]^{1/2}$$

Now consider an additional observation (X_o, Y_o). For given X_o, it is possible to derive an interval that possesses a 90-percent probability of capturing Y_o. This interval is given by the formula

$$\hat{a} + \hat{b}X_o \pm t_{n-2,.95}\, S\left[1 + \frac{1}{n} + \frac{(X_o - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\right]^{1/2}$$

The estimates of the intercept and slope may also be used to correct the model for linear trend. To accomplish this, we replace the model prediction Y by the corrected prediction $(Y - \hat{a})/\hat{b}$.

To supplement the material presented herein, the reader is referred to the text by Miller.[1]

PEARSON'S CORRELATION COEFFICIENT

Direct correlation between observed and predicted concentrations exists when a rise or fall in predicted concentrations is associated with a corresponding linear rise or fall in observed concentration.

The degree of correlation between predicted and observed concentrations can often be visually assessed through the use of a scatterplot, as illustrated in Figure D-2.

In the situations shown in Figure D-2, the more closely the plotted concentrations can be fitted by a line, the more highly correlated are predicted and observed concentration. A correlation of +1 denotes that all the plotted points fall along a line; a correlation close to zero denotes that the data scatters widely.

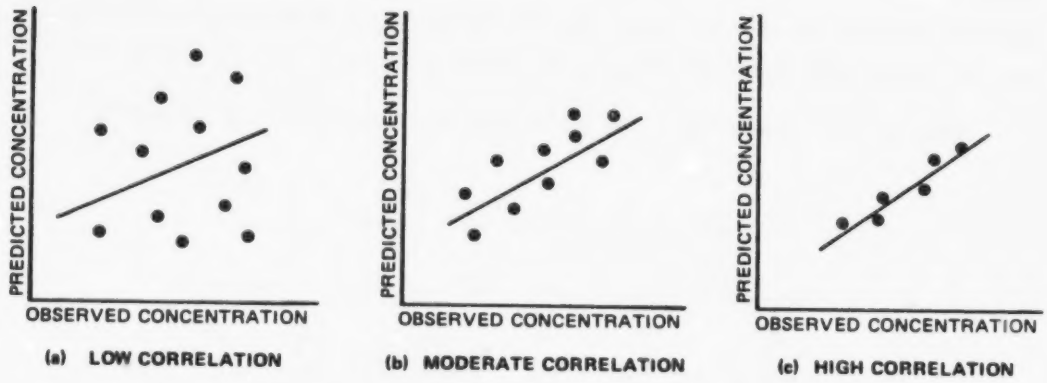

Figure D-2. Visual assessment of correlation.

Two anomalous situations may also occur. In the first, the correlation coefficient is negative. The line that best fits the scatterplot has a negative slope. A correlation of -1 denotes that all the plotted points fall along such a line. In the second, the predicted and observed concentration scatterplot shows a definite nonlinear relationship. As illustrated in Figure D-3, the correlation will be low in such cases unless the data can also be fit well by a line.

When the relationship between predicted and observed concentration is linear, the square of the correlation coefficient may be interpreted as the percentage of the variance of the predicted concentrations

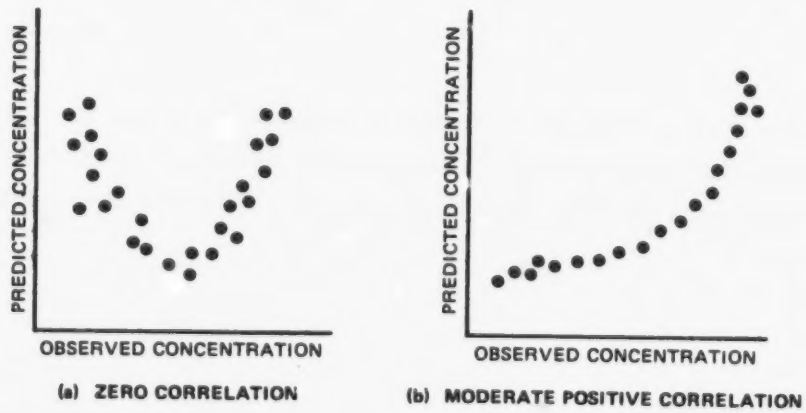

Figure D-3. Nonlinear scatterplot relationships.

"explained" or "accounted for" by the observed observations. The converse interpretation--that the squared correlation coefficient represents the percentage of the variance of the observed concentrations explained by the predicted observations--also holds.

Pearson's formula for the correlation coefficient is given by

$$r = \frac{n\sum_{i=1}^{n}(OC_i)(PC_i) - \left(\sum_{i=1}^{n}OC_i\right)\left(\sum_{i=1}^{n}PC_i\right)}{\sqrt{n\sum_{i=1}^{n}OC_i^2 - \left(\sum_{i=1}^{n}OC_i\right)^2}\ \sqrt{n\sum_{i=1}^{n}PC_i^2 - \left(\sum_{i=1}^{n}PC_i\right)^2}}$$

In discussing the correlation coefficient, we have been concerned with the correlation exhibited in the sample. We now consider the correlation coefficient in the population from which the sample has been drawn. Due to sampling variability, Pearson's correlation coefficient for the sample will not usually equal the correlation coefficient in the population. With a large sample, however, the two correlation coefficients will be close.

The correlation coefficient for the population is usually denoted "ρ." It can be shown that if the observed and predicted concentrations are jointly normally distributed, there is a 95-percent probability that ρ is contained in the interval

$$\tanh\left(\tanh^{-1}r - \frac{1.96}{\sqrt{n-3}}\right) \leq \rho \leq \tanh\left(\tanh^{-1}r + \frac{1.96}{\sqrt{n-3}}\right)$$

This interval is called a 95-percent confidence interval for the population correlation coefficient. The texts by Anderson[2] and Wonnacott[3] provide more detailed information on the derivation of the confidence interval.

TEMPORAL AND SPATIAL CORRELATION COEFFICIENTS

As described by Nappo,[4] an accurate air-prediction model must reproduce at each monitoring station the observed time-varying pollution concentration, and reproduce at each monitoring time the observed space-varying pollution pattern. A measure of the agreement of the predicted and observed temporal trends of air pollution over the whole network of monitoring stations is given by the temporal correlation coefficient $\overline{R(t)}^{s}$, the average over stations (spatial average) of the temporal correlation coefficient formed at each station. A measure of the agreement of the predicted and observed spatial trends of air pollution over a prediction period is given by the spatial correlation coefficient $\overline{R(s)}^{t}$, the average over time of the correlation coefficient between the predicted and observed patterns of the concentration isopleths at each monitoring time.

Let the collection of observed and predicted concentrations, $O_{s,t}$ and $P_{s,t}$, be indexed by the monitoring station $s = 1,2,\ldots,S$ and by the monitoring time $t = 1,2,\ldots,T$. It is not necessary for the stations or the times to be contiguous.

The time-averaged, station-specific, predicted and observed concentrations are given by

$$\overline{P_{s,t}}^{t} = \frac{1}{T}\sum_{t=1}^{T} P_{s,t}$$

$$\overline{O_{s,t}}^{t} = \frac{1}{T}\sum_{t=1}^{T} O_{s,t}$$

The station-specific temporal correlation coefficient is given by the usual formula,

$$R_s(t) = \frac{\sum_{t=1}^{T}(P_{s,t} - \overline{P_{s,t}}^{t})(O_{s,t} - \overline{O_{s,t}}^{t})}{\sqrt{\sum_{t=1}^{T}(P_{s,t} - \overline{P_{s,t}}^{t})^2}\sqrt{\sum_{t=1}^{T}(O_{s,t} - \overline{O_{s,t}}^{t})^2}}$$

The spatial average of the temporal correlation coefficient is given by

$$\overline{R(t)}^{s} = \frac{1}{S}\sum_{s=1}^{S} R_s(t)$$

Under the assumption that the $R_s(t)$ are not related to each other (i.e., statistically independent) and have identical frequency distributions with mean $E[R_s(t)] = R(t)$, a 95-percent confidence interval for $R(t)$ is given by

$$\tanh\left[\frac{1}{S}\sum_{s=1}^{S}\tanh^{-1}R_s(t) - 1.96\frac{1}{\sqrt{S(T-3)}}\right] \leq R(t)$$

$$\leq \tanh\left[\frac{1}{S}\sum_{s=1}^{S}\tanh^{-1}R_s(t) + 1.96\frac{1}{\sqrt{S(T-3)}}\right]$$

The station-averaged, time-specific, predicted and observed concentrations are given by

$$\overline{P_{s,t}}^{s} = \frac{1}{S}\sum_{s=1}^{S} P_{s,t}$$

$$\overline{O_{s,t}}^{s} = \frac{1}{S}\sum_{s=1}^{S} O_{s,t}$$

The time-specific station correlation coefficient is given by the usual formula,

$$R_t(s) = \frac{\sum_{s=1}^{S} (P_{s,t} - \overline{P_{s,t}}^{s})(O_{s,t} - \overline{O_{s,t}}^{s})}{\sqrt{\sum_{s=1}^{S} (P_{s,t} - \overline{P_{s,t}}^{s})^2} \sqrt{\sum_{s=1}^{S} (O_{s,t} - \overline{O_{s,t}}^{s})^2}}$$

The temporal average of the time-specific spatial correlation coefficient is given by

$$\overline{R(s)}^{t} = \frac{1}{T} \sum_{t=1}^{T} R_t(s)$$

Under the assumption that the $R_t(s)$ are independent and identically distributed with mean $E[R_t(s)] = R(s)$, a 95-percent confidence interval for $R(s)$ can be formulated as

$$\tanh\left[\frac{1}{T}\sum_{t=1}^{T}\tanh^{-1}R_t(s) - 1.96\frac{1}{\sqrt{T(S-3)}}\right] \leq R(s)$$

$$\leq \tanh\left[\frac{1}{T}\sum_{t=1}^{T}\tanh^{-1}R_t(s) - 1.96\frac{1}{\sqrt{T(S-3)}}\right]$$

CHI-SQUARE GOODNESS-OF-FIT TEST

The chi-square goodness-of-fit test can be used to assess whether the distribution of predicted concentrations deviates significantly from the distribution of observed concentrations, and to graphically depict such departure. Differences between the observed and predicted concentration distributions, such as different variances or shapes, may be

identified from appropriate data displays. For more detailed information on this test, the reader is referred to the text by Hogg and Craig.[5]

The observed concentrations $OC_1,\ldots,OC_n$ are used to define cell boundaries for air-pollution concentration with an approximately equal number of observed concentrations in each cell. If possible, each cell should have approximately ten observations. By E_j, denote the number of observed concentrations in the j-th cell, where there are J cells, and $1 \leq j \leq J$. By P_j, denote the number of predicted concentrations of the j-th cell. If the model is accurate, the E_j and the P_j should be approximately equal.

Under the assumption that the E_j are the expected number of predicted concentrations in the j-th cell, the statistic

$$\chi^2 = \sum_{j=1}^{J} \frac{(P_j - E_j)^2}{E_j}$$

is distributed as a chi-square random variable with J - 1 degrees of freedom.

Larger values of the chi-square statistic than might reasonably be anticipated from a chi-square random variable are indicative of poor model fit.[3] The significance of the chi-square statistic may be obtained from a chi-square distribution table. It is usual practice to report the value of the chi-square statistic (χ^2) as well as its significance (obtained from appropriate tables). However, both these values are very sensitive to sample size. We suggest also reporting χ^2/n, which may be interpreted as the sum over the cells of the percentage deviation from the expected cell proportions.

INTERSTATION ERROR CORRELATION TEST

Any correlation between the errors in concentration prediction of two receptor stations is indicative of a systematic bias in the predictor model. The significance of these error correlations may be tested and confidence bounds placed on the error correlations.

A correlation coefficient that is approximately zero between the errors in concentration prediction of two receptor sites is a highly desirable situation. It indicates that no important condition "linking" the two receptor sites exists that might account for underprediction or overprediction.

If the correlation coefficient is high (near 1.0), it indicates that whenever the model overestimates concentrations at one receptor site it tends to do so at the other, and similarly for underestimation. The presence of a factor influencing both observed concentrations but unaccounted for by the model is suggested.

If the correlation coefficient is low (near -1.0), it indicates that whenever the model overestimates concentrations at one receptor site, it tends to underestimate concentrations at another, and, conversely, whenever the model underestimates at one receptor site it tends to overestimate at the other. The presence of a factor affecting both observed concentrations in opposite directions, but unaccounted for by the model, is suggested.

For two stations j and k, let the set of observed and predicted concentrations be given by (OC_{j1},PC_{j1}), ..., (OC_{jn},PC_{jn}) and (OC_{k1},PC_{k1}), ..., (OC_{kn},PC_{kn}). The differences between the observed and predicted concentrations will be denoted

$$DC_{jt} = OC_{jt} - PC_{jt}$$

where j (or k) is the station identifier and t is the time identifier.

Let r be the correlation coefficient between $(DC_{ji}:i = 1, \ldots, n)$ and $(DC_{ki}:i = 1, \ldots, n)$. Under the assumption that DC_j and DC_k follow

a bivariate normal distribution with correlation coefficient ρ, a 95-percent confidence interval[2] is given by

$$\tanh\left(\tanh^{-1}r - \frac{1.96}{\sqrt{n-3}}\right) \leq \rho \leq \tanh\left(\tanh^{-1}r + \frac{1.96}{\sqrt{n-3}}\right) .$$

If this confidence interval fails to capture the value zero, then the correlation coefficient ρ is significantly different from zero.

MULTIPLE REGRESSION OF ERROR RESIDUALS

Multiple regression[6] is a technique for uncovering linear relationships between the residuals (observed concentrations less predicted concentrations) and the air-pollution/meteorological variables. Figure D-4 illustrates a hypothetical plot of residuals versus ambient air temperature. In this case, the prediction model tends to underestimate concentrations at high temperatures and overestimate concentrations at low temperatures. The elimination of this linear trend through model

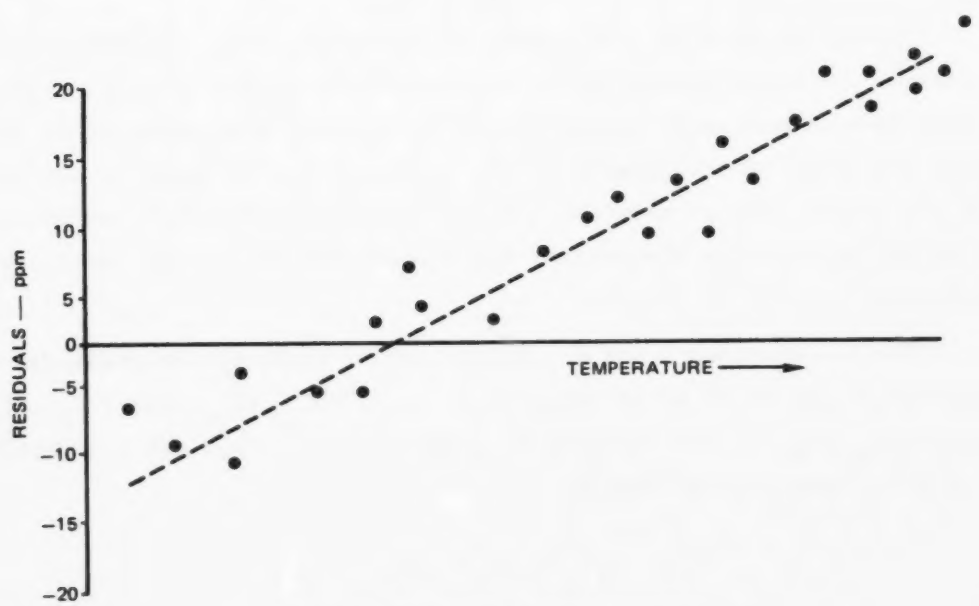

Figure D-4. Sample residual versus parameter plot.

adjustment will improve the accuracy of the prediction. Multiple linear regression can be used to uncover these linear relationships when there are many air pollution/meteorological variables.

To perform a multiple linear regression, we must have data consisting of a set of n paired measurements of observed concentrations, predicted concentrations, and air-pollution/meteorological variables, $(OC_1,PC_1,X_{11},\ldots,X_{1p})$, ..., $(OC_n,PC_n,X_{n1},\ldots,X_{np})$. The air-pollution/meteorological variables can be interval, ordinal, or nominal 0-1 variables. They may even include predicted and observed concentrations.

The residual Z_k is the difference between the k-th observed and predicted observation

$$Z_k = OC_k - PC_k \qquad (1 \leq k \leq n)$$

To uncover linear relationships between the residual and the air-pollution/meteorological variables, we fit the following model:

$$Z_k = \mu + \sum_{j=1}^{P} \beta_j X_{kj} + e_k \qquad (1 \leq k \leq n)$$

Here the e_k are independent variables following a normal $(0,\sigma^2)$ distribution. The usual least-square procedures are used to estimate μ and β coefficients.

The estimate of μ is the amount by which the model, on the average, underestimates or overestimates the observed concentration. The estimates of the β_j are the slopes of the linear relationship of the residuals to the air-pollution/meteorological variables.

The statistical significance of the estimates of the β_j may be determined by using the F-test. If the F-test is significant at the 5-percent level, then consideration should be given to using the regression results to modify the model.

The multiple regression should be performed in a stepwise fashion in order to isolate the most important subset of the air-pollution/meteorological variables. These variables can be identified as providing the largest increments to the proportion of the explained variance in the residuals. Efforts at model modification should be focused on these variables.

ANALYSIS OF THE RESIDUAL TIME SERIES

Statistical techniques, as described by Box and Jenkins[7] and Jenkins and Watts,[8] may be used to examine the residual time series (observed concentrations less predicted concentrations) to determine if components other than white noise are present.

Due to the serial and cyclical time dependence of meteorological conditions and pollutant emissions, both the observed-concentrations series and predicted-concentrations series will also exhibit time dependencies. To the extent, however, that the predicted time series agrees with the observed concentrations, the residual time series (observed concentrations less predicted concentrations) will consist primarily of white noise.

If there is a serial or cyclical component in the residual time series, then either the prediction model is failing to account for a time-dependent relationship or it is inducing an extraneous time-dependent relationship. In either event, a fault in the prediction model is indicated.

To perform the time-series analysis, we need data consisting of a set of n paired measurements of observed concentrations and predicted concentrations equally spaced in time: (OC_1,PC_1), (OC_2PC_2), ..., (OC_n,PC_n). These data are used to form the residual time series

$$r_t = OC_t - PC_t \qquad (t = 1, \ldots, n)$$

The residual time series is then investigated for serial and cyclical departures from white noise.

The presence of serial departure from white noise may be detected by using the autocorrelation function. The autocorrelation function of the residual time series is given by

$$A(k) = \frac{\frac{1}{n-k}\sum_{t=1}^{n-k}(r_t - \bar{r})(r_{t+k} - \bar{r})}{\frac{1}{n}\sum_{t=1}^{n}(r_t - \bar{r})^2}$$

where

$$\bar{r} = \frac{1}{n}\sum_{t=1}^{n} r_t$$

and k indexes the lag of the serial dependence (i.e., daily, weekly, bimonthly, monthly).

If the r_t series is a white-noise process, then any A(k) is asymptotically normally distributed with mean zero and variance 1/n, and should lie inside the 95-percent confidence limits of $\pm 1.96/\sqrt{n}$. If an A(k) value is outside these limits, the possibility of a time dependency should be closely investigated.

Figure D-5 illustrates the confidence bands for the autocorrelation function. Two autocorrelations (at lag one and seven days) fall outside the confidence bands. These autocorrelations indicate that if the model overestimates on one day, it tends to underestimate the next, and if the model overestimates on one day of the week, it tends to overestimate on the same day of the next week.

Finally, it should be noted that of every twenty or so autocorrelations tested, about one will fall outside the bands by mere chance. Therefore, each autocorrelation falling outside the bands needs to be more closely investigated (either analytically or by means of another time series) to establish that the observed result is not a chance artifact.

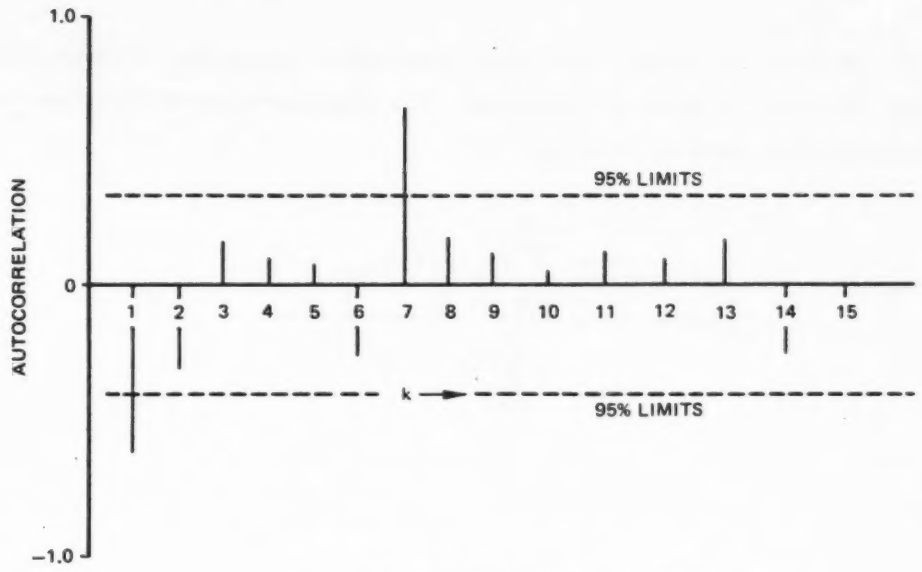

Figure D-5. Sample autocorrelation plot.

The presence of cyclical deviation from white noise may be detected by using the normalized cumulative periodogram. The normalized cumulative periodogram is given by

$$C(f_j) = \frac{\sum_{i=1}^{j} I(f_i)}{\sum_{t=1}^{n} (r_t - \bar{r})^2}$$

where

$$I(f_i) = \frac{2}{n}\left[\left(\sum_{t=1}^{n} r_t \cos 2\pi f_i t\right)^2 + \left(\sum_{t=1}^{n} r_t \sin 2\pi f_i t\right)^2\right]$$

is the periodogram of the residual time series and where $f_i = i/n$ is the frequency.

For a white-noise series, the plot of $C(f_i)$ against f_i would be close to a straight line joining the points (0,0) and 0.5,1). Ninety-five-percent confidence limits can be drawn at a distance of $\pm 1.96/\sqrt{n}$ above and below this straight line, as shown in Figure D-6. Note that $C(f_i)$ crosses the confidence limits. This indicates that there are cyclical deviations from white noise. The largest jumps in $C(f_i)$ occur in a pattern of about seven days and therefore the model should be investigated to determine why it tends to overestimate and underestimate observed concentrations with this frequency.

To use these two time-series analysis techniques, we must have a data set consisting of equally spaced readings in time. These readings could be hourly, daily, weekly, monthly, etc. It is recommended that at least 50 consecutive readings be included in each autocorrelation and

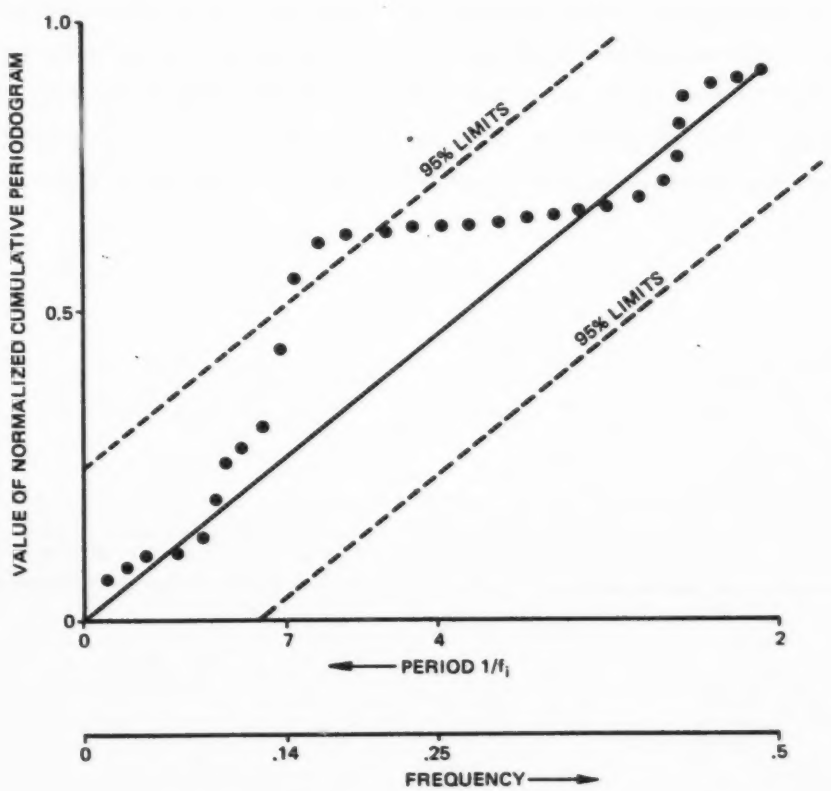

Figure D-6. Sample plot of cumulative periodogram.

periodogram calculation. It is not necessary that all values be consecutive since the computational formulas may be adjusted for occasional missing values.

RUNS TEST

The Wald-Wolfowitz runs test[9] is a method for determining whether the prediction model overestimates or underestimates the observed concentration for long periods of time.

Consider a plot of the amount of over- or under-prediction of SO_2 concentrations, as shown in Figure D-7. The dashed line represents the median overprediction (nine observations are below the median and nine above). Adjusting for the median overprediction, we obtain Figure D-8. Note that the plotted points tend to stay above or below the median for extended periods of time. This indicates that there is a nonrandom component of prediction error present (we have already subtracted out the constant overprediction component). If only random noise were present, the plotted errors in prediction would fluctuate more erratically around the median. The presence of only four runs (sequence above or below the median of one or more points) indicates the presence of a nonrandom component.

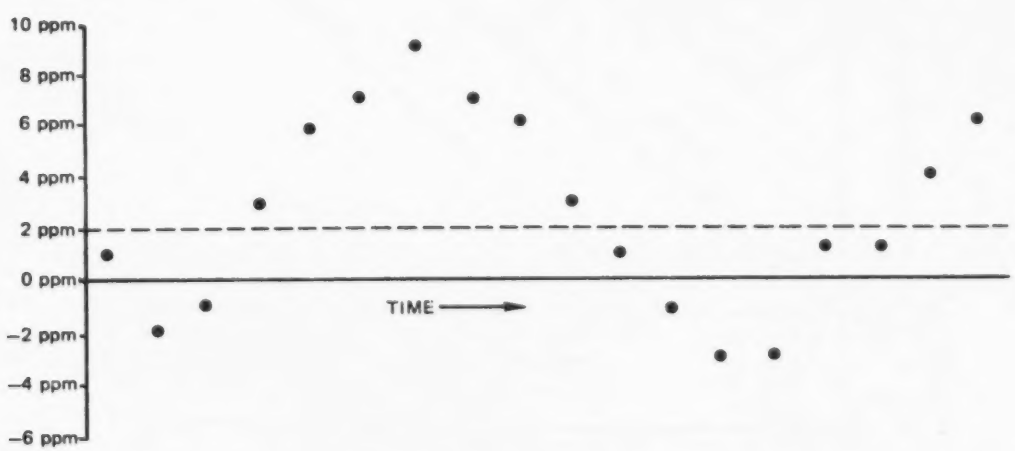

Figure D-7. Sample plot of residuals versus time.

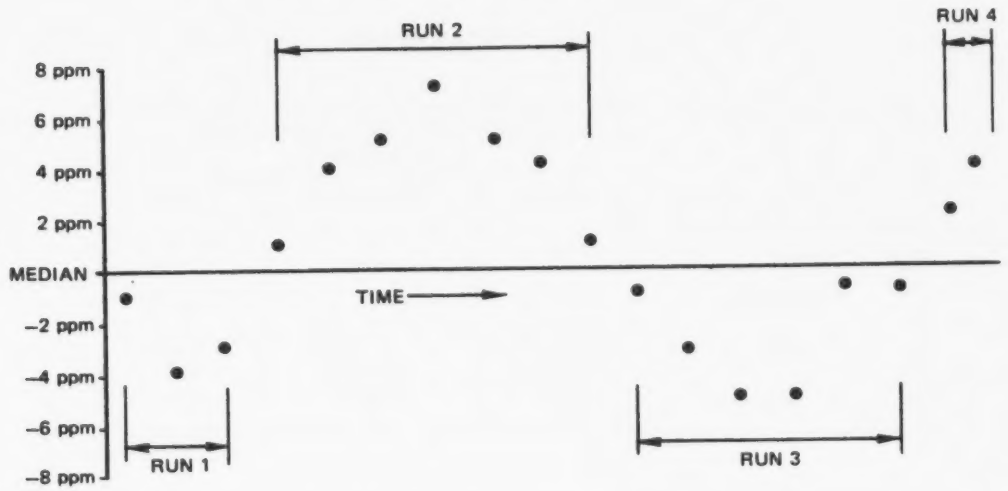

Figure D-8. Adjusted sample plot of residuals versus time.

If the Wald-Wolfowitz runs test is used, the observed and predicted concentration readings should be consecutively, and approximately equally, spaced in time. We can conclude that there is a nonrandom component in the sequence if there are too few runs. The minimum allowable number of runs over n time periods is

$$\frac{n + 2}{2} - \frac{1.65}{2}\sqrt{n - 2}$$

Fewer runs than this indicates, with 95-percent confidence, the presence of a nonrandom component of error. This formula is fairly accurate for over 40 data points. Otherwise, appropriate tables must be used.

POST-AGREEMENT AND PREFIGURANCE TABLES

The post-agreement and prefigurance tables are two methods that are useful in assessing the accuracy of a forecast model.[10] The post-agreement table (Table D-1) allows an assessment of the extent to which subsequent observations confirm the prediction of a certain pollution level. The prefigurance table (Table D-2) allows an assessment of the extent to which the forecasts give advance notice of a certain pollution level.

TABLE D-1. POST-AGREEMENT

Observed Level	Forecast Pollution Level (Percent)		
	Heavy	Moderate	Light
Heavy	57.1	21.1	2.5
Moderate	19.0	53.3	10.0
Light	23.9	25.6	87.5
Total	100	100	100

TABLE D-2. PREFIGURANCE

Observed Level	Forecast Pollution Level (Percent)			
	Heavy	Moderate	Light	Total
Heavy	54.5	40.9	4.6	100
Moderate	14.3	71.4	14.3	100
Light	10.0	20.0	70.0	100

In Tables D-1 and D-2, three levels (heavy, moderate, and light) have been used. In general, the number of levels and their specification in terms of pollutant concentrations are determined by the user.

The post-agreement table will be important in deciding whether a forecast model is accurate enough to call an alert when heavy pollution is forecast. The prefigurance table will be important in deciding whether a forecast model is correctly forecasting a large enough percentage of heavy-pollution days.

QUASI-LOGISTIC MODELING OF THE CORRELATION COEFFICIENT

The Pearson's correlation coefficient between observed and predicted concentrations may be computed for many different subsets of the data (such as days with a particular mixing height, wind speed, ambient air temperature, etc.). Some subsets will be associated with high correlation coefficients and others with low correlation coefficients. Quasi-logistic modeling of the correlation coefficient is a method of assessing which of the subset-defining variables (mixing height, wind speed, etc.) are contributing to or reducing the correlation coefficient. That is, quasi-logistic modeling allows a determination of the conditions under which the observed and predicted concentrations are well or poorly correlated.

Suppose we are interested in the effect of m subset-defining variables on the correlation coefficient. These "independent" variables will be denoted $X_1, \ldots, X_m$. In the spirit of the logistic model, we model the true correlation coefficient as

$$\rho = \frac{\exp\left(A_o + \sum_{i=1}^{m} A_i X_i\right) - 1}{\exp\left(A_o + \sum_{i=1}^{m} A_i X_i\right) + 1}$$

Here the $A_o, A_1, \ldots, A$ are coefficients. Of course, we have not yet specified how to estimate the coefficients nor how to assess whether the model fits the correlation coefficient. But we should note some features of this model. If A_o is large in comparison with the other A_i (even when multiplied by the X_i), then the true correlation coefficient is essentially constant over all subsets of the data. On the other hand, whenever any $A_i X_i$ become large and positive, it follows that the correlation coefficient is especially large due to the contribution of the i-th subset-defining variable.

For any particular subset of the data, we can compute Pearson's correlation coefficient r. Under the usual bivariate normality assumptions for predicted and observed concentrations, the quantity

$$Z = \ln \frac{1 + r}{1 - r}$$

is approximately normally distributed with variance equal to $4/(n - 3)$. Here n refers to the number of data points in the subset used to calculate r. The mean for Z is

$$\ln \frac{1 + \rho}{1 - \rho}$$

which under our model for ρ reduces to

$$A_o + \sum_{i=1}^{m} A_i X_i$$

A weighted least-squares regression of Z on the X_i may be used to estimate the A_o, A_i, ..., A_m coefficients. Many packaged computer routines, such as SPSS,[11] exist for performing this regression.

The overall fit of the model may be assessed through the use of the multiple R^2 statistic and plots of the residuals (calculated Z values less estimated Z values). The multiple R^2 statistic (and its associated F value) indicate the percentage reduction in the variance of Z accounted for by the regression. The residuals pinpoint those data subsets for which the model is particularly ill-fitting, perhaps indicating the presence of "interactions" among variables.

The partial R^2 statistics (and their associated F values) may be used to single out the regression coefficients that are particularly effective in accounting for the variability in the correlation coefficient. This may allow a simplification of the model by reducing the number of coefficients.